Description

Method, computer program product with program code means and computer program product for analysis of a regulatory genetic network of a cell

The invention relates to an analysis of a regulatory genetic network of a cell using a statistical method.

Fundamentals of a regulatory genetic network of a cell are known from [1]. Such a regulatory genetic network should be taken in this document to mean in particular regulatory interactions between genes of a cell.

A genome, i.e. the human genetic substance, is estimated to comprise 20,000 to 40,000 genes, of which a biologically specified number in each case- depending on a specialization of a cell - are present in the cell in the form of a DNA or a part of a DNA.

A not necessarily contiguous section of this DNA containing the genetic code for a protein or also for a group of proteins or for creating a protein or a group of proteins is designated as a gene here. Overall the genes contain a genetic code for around a million proteins.

An interplay or the interactions between the genes as well as with the proteins represents the most important part of a machinery (regulatory genetic network) which underlies the development of a human body from a fertilized egg cell as well as all bodily functions.

It is also known from [1] that so-called gene expression rates which form a gene expression pattern supply a description or representation of a regulatory genetic network or of a current status of the regulatory genetic network.

In simple terms or expressed more clearly the gene expression pattern of a cell thus represents a state of the regulatory genetic network of this cell.

It is further known that by using high-throughput gene expression measurements (microarray data) these gene expression rates can be measured. The microarray data in its turn describes snapshots of the gene expression pattern.

Many illnesses and malfunctions of the body are attributable to disturbances in the regulatory genetic network which is reflected by greatly changed gene expression behavior (gene expression rates) or a changed gene expression pattern of a cell.

An understanding of the regulatory genetic network thus represents an important step on the path to a characterization of the understanding of genetic mechanisms as well as consequently of identification of what are known as dominant or malfunction-initiating genes underlying the illnesses or malfunctions.

In cancer research for example suppressing genes can play a key role in the identification of growths and tumors, the knowledge of new potential oncogenes and their interactions with other genes can be a contribution to discovering the basic principles (of cancers) which determine how normal cells change into malignant cancer cells.

Furthermore a quantitative understanding of the regulatory genetic network of a cell is necessary for developing improved medicaments and therapies for fighting genetic diseases.

Thus a number of medicaments act as agonists or antagonists of specific target proteins, i.e. they strengthen or weaken the function of a protein with corresponding effect on the

regulatory genetic network with the aim of bringing this back into a normal function mode.

A description of a regulatory genetic network of a cell using a statistical method, a causal network is known from [2].

A causal network, a Bayesian network, is known from [3].

Bayesian networks

A Bayesian network *B* is a specific type of presentation of a common multivariate probability density function (WDF) of a set of variables X by a graphical model which consists of two parts.

It is defined by a directed acyclic graph, *DAG*) *G* - of the first component, in which each node i = 1, ...,n corresponds to a random variable X_i.

The connectors between the nodes represent statistical dependencies and can be interpreted as causal relationships between them. The second component of the Bayesian network is the set of conditional WDFs $P(X_i | Pa_i, \theta, G)$, which are parameterized by means of a vector θ.

These conditional WDFs specify the type of dependencies *of the individual variablesi of the set of its parents* $_{Pa}$i. Thus the common WDF can be broken down into the product form

(1) $$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i | Pa_i, \theta, G)$$

The *DAG* of a Bayesian network uniquely describes the conditional dependency and independency relationships between a set of variables, but by contrast a given statistical

structure of the WDF does not result in any unique DAG.

Instead it can be shown that two *DAGs* describe one and the same WDF, if and only if they feature the same set of connectors and the same set of "colliders", with a collider being a constellation in which at least two directed connectors lead to the same node.

The object of the invention is to specify a method which allows an analysis of a regulatory genetic network of a cell, for example represented by at least one gene expression pattern of the cell.

A further object of the invention is to specify a method which enables a defective gene to be identified, for example a cancer or tumor gene, in the regulatory genetic network of a cell.

Further the invention is designed to allow a simulation and/or an analysis of an effect of a medicament on the regulatory genetic network of a cell.

This object is achieved by the method, the computer program product with program code means and the computer program product for analysis of a regulatory genetic network of a cell with the features according to the relevant independent patent claim.

In the basic method for analysis of a regulatory genetic network of a cell a causal network is used,

- said causal network describing the regulatory genetic network of the cell such that nodes of the causal network represent genes of the regulatory genetic network and connectors of the causal network represent regulatory interactions between the genes of the regulatory genetic network

In the analysis method a gene expression rate is now specified for a selected gene of the regulatory genetic network. Using the causal network a resulting gene expression pattern is generated for the predetermined gene expression rate for the regulatory genetic network. The resulting gene expression pattern generated is subsequently compared with a predetermined gene expression pattern of the regulatory genetic network.

The computer program product with program code means is set up to execute all the steps in accordance with the inventive method when the program is executed on a computer.

The computer program product with program code means stored in machine-readable form on a data medium is set up to execute all the steps in accordance with the method in accordance with the invention when the program is run on a computer.

The arrangement and also the computer program product with program code means set up to execute all steps in accordance with the inventive method when the program is run on a computer, as well as the computer program product with program code means stored on a machine-readable medium, set up to execute all steps in accordance with the inventive method when the program is executed on a computer are especially suited to execute the method in accordance with the invention or of one of its further developments listed below.

A probabilistic semantic of a causal network, such as of a Bayesian network, is very well suited to analysis of gene expression rates, given for example in the form of microarray data, since it is adapted to the stochastic nature both of biological processes and also to experiments susceptible to noise.

Furthermore, viewed in illustrative terms, an effect of an

expression state of specific genes on a global gene expression pattern (inverse modeling) is estimated, in that a resulting gene expression pattern is analyzed.

Preferred developments of the invention are produced by the dependent claims.

The developments described below relate to both the method and to the configuration.

The invention and the developments described below can be implemented both in software and also in hardware, for example by using a specific electrical circuit.

Further the realization of the invention or of a development described below is possible through a computer-readable storage medium on which a computer program product with program code means is stored which executes the invention or development.

Also the invention or any development of it described below can be realized by a computer program product which features a storage medium on which a computer program product with program code means is stored which executes the invention or development.

With a further development the selected gene is selected using the causal network by means of a dependency analysis.

The gene expression rate of the selected gene can also be predetermined such that the predetermined gene expression rate of the selected gene reflects an assumption of a gene defect.

A Bayesian network can be used as the causal network.

The causal network can also be of a type DAG (Directed Acylic Graph).

Furthermore the generated resulting and/or the predetermined gene expression pattern can represent discrete gene states, with the represented discrete gene states being able to be a an overexpressed, a normal or an underexpressed gene state.

In a further development the generated resulting gene expression pattern can be compared with the predetermined gene expression pattern using a static method and/or of a statistical code, especially a measure of distance.

There can also be provision for the causal network to be trained using gene expression patterns, with the nodes and the connectors of the causal network being adapted.

Furthermore it is expedient for the gene expression patterns, especially the predetermined gene expression pattern and/or the gene expression patterns for training, to be determined using a DNA microarray technique.

In one embodiment the predetermined gene expression pattern and/or the gene expression pattern for training is a gene expression pattern of a genetic regulatory network of a diseased cell.

Here for example the diseased cell can be a cancer cell, especially a oncocell with ALL (Acute Lymphoblastic Leukemia).

Furthermore the diseased cell can feature an oncogene, especially an ALL oncogene.

Also for a plurality of selected genes of the regulatory genetic network one gene expression can be predetermined in each case, a plurality of resulting gene expression patterns generated and/or a plurality of comparisons undertaken.

In a further development the generation of the plurality of resulting gene expression patterns is performed iteratively.

Furthermore the inventive procedure or development is particularly suitable for identifying a dominant gene and/or a degenerated/mutated/diseased gene/oncogene/tumor-suppressor gene.

It is also suitable for identifying a tumor cell, for example in connection with cancer detection.

Further the inventive method is especially suited to analyzing the causes of an abnormal gene expression pattern/ gene expression rate.

It can also be used for a simulation and/or analysis of the effects of a medicament.

The figures show an exemplary embodiment of the invention which is explained in more detail below.

The figures show

Figure 1 a digram of a procedure for investigating genetically-related causes of illness through Bayesian inverse modelling using a cancer as an example;

Figure 2 a digram with an algorithm for creating a data set of N samples in accordance with an exemplary embodiment;

Figure 3 a digram of a procedure for creating data sets, which reflect an effect of different observations in accordance with an exemplary embodiment;

Figures 4a and b digrams which show that data obtained by sampling show subtype characteristic expression patterns as also in an original data set;

Figure 5 a diagram which shows graphically a probability of each subtype under a condition which is overexpressed

on a gene, for all 271 genes;

Figure 6 a diagram of a graph structure of a causal network, which represents a regulatory genetic network.

Exemplary embodiment Investigation of genetically-related causes of diseases using Bayesian inverse modelling using a cancer as an example (espec. Fig.1)

Overview of the Bayesian Inverse Modelling (BIM) procedure

In many areas of empirical research the desire is to reach conclusions from the observation of trial results about the underlying principle and its causes - the relationship between "cause" and "effect".

For example in cancer research the underlying principle is studded which causes a normal cell to transform it into a malignant, rapidly growing cancer cell.

The effect of the various types of cancer is known, e.g. the general appearance of a cancer cell compared to a normal cell, measured with the aid of microarray chips.

By contrast the cause of its origination is largely unknown.

On the basis of the understanding that cancer is a genetic illness and that it is attributable to a deviation in the behavior of cells, the research is concentrating on discovering the genetic principles which are responsible for the development of the cancer.

An important task in this environment is to identify genes which can play a role in tumor genesis, such as for example growth and tumor-suppressing genes.

A procedure is described below with which it is possible to identify genes which are a potential cause of tumor genesis.

One element of the procedure is a statistical method, in this case a Bayesian network [3] (see above and subsequent associated embodiments for more details), which is learnt [2] from a microarray data set [1] (see "Structural learning" below) (cf. **Fig. 1**).

In this case it is assumed that the set of the measured gene expression vectors X belong to a basic totality with a highly-dimensional multivariate probability density function which is modelled with the aid of Bayesian network with adaptive network structure.

The relationships between the variables, namely the conditional dependences and independences, are represented by means of a Directed Acyclic Graph (DAG) G .

The probabilistic semantic of the Bayesian network is very well suited to the analysis of microarray data since it is adapted to the stochastic nature both of the biological processes and also of the experiments susceptible to noise.

In the procedure described below the learnt Bayesian network will be used as a generative model for taking samples of artificial microarray data sets which supplies the learned conditional probability density distributions (cf. **Fig.1, step 110 - 130**).

Furthermore the effect of the expression state of specific genes on the global gene expression pattern (inverse modelling) is estimated, in that a resulting data set is analyzed (cf. **Fig.1, step 110 - 130**).

In the procedure described below each gene is also assigned its probability, with which it is the cause of these cell states.

To this end these data sets are compared with data obtained

from microarray investigations of various known cell states (cf. **Fig.1, step 130**).

Seen in general terms, the procedure does not concentrate explicitly on the structures of the network, but rather on the probability distribution which is derived from the learnt Bayesian network.

Finally the procedure is applied to microarray data of different subtypes of pediatric acute lymphoblastic leukemia (ALL) of Yeoh et al. [4].

The comparison of the artificial data with expression patterns of specific cancer subtypes enables a measure of probability of the illness-causing behavior of each gene (cf. **Fig.1, step 130**) to be obtained.

Results of the applied procedure show that, in connection with Bayesian Inverse Modelling (BIM) this allows the effect of pathogenetiically modified expression levels on the global gene expression pattern to be predicted, in which case already known oncogenes as well as potential new ones are found.

Bayesian networks

The basic principles of Bayesian networks [3] have already been described above.

In the case of the modelling of a regulatory genetic network by a Bayesian network genes or their corresponding proteins are symbolized by nodes.

Regulation mechanisms are described by connectors between two nodes, which can be interpreted in a causal manner.

The quality of the regulation is encoded in the conditional probability distribution of the gene involved for given

regulators of the same.

Structural learning

The process of structural learning can be described as follows:

Let $D = \{d^1, d^2, \ldots, d^N\}$ be a data set of N independent observation, with each data point being an n-dimensional vector with components $d^1 = \{d^1_1\ d^1_2, \ldots, d^1_N\}$. For a given D the structure G of the Bayesian network is to be found which best corresponds to D, i.e. which maximizes the Bayes-Score,

(2) $$S(G|D) = \frac{P(D|G)P(G)}{P(D)}$$

with $P(D|G)$ the being the peripheral probability, $P(G)$ the apriori probability of the structures and $P(D)$ the evidence.

Since both the apriori probabilityand also the evidence are unknown, the problem is reduced to determining the structures with the best peripheral probability corresponding to the data (Heckerman et al. [5]).

If the data set D consists of N microarray experiments, e.g. of cell samples of different patients, each data vector $\{d^1_1, d^1_2, \ldots, d^1_n\}$ represents the expression profile of n genes in a microarray experiment.

A Bayesian network learnt from such data encodes the probability distribution of n genes, which were obtained from these N microarray experiments.

Bayesian Inverse Modelling (BIM)

Generative model

A learnt (see notes above about "structural learning") Bayesian network B represents a density estimation function which reflects the probability distribution of the data set D, on the basis of which it was learnt, with the aid of the set of conditional WDFs.

This means that it can be used as a generative model for creating a data set D_B which reflects the density distribution obtained from D.

Fig. 2 shows an algorithm 200 for creating a data set of N samples from B.

The first step 210 of the algorithm 200 consists of arranging all variables such that the parents (parent nodes) Pa_i are instantiated before X_i.

Subsequently the variables corresponding to the arrangement are selected and instantiated with a value 220.

The value of each variable is selected with the probability $P(\text{state}|Pa_i)$. This step is repeated 230, until N samples are created.

Probabilistic interference

A significant problem in Bayesian networks is the evidence propagation, meaning the determination of the aposteriori distribution $P(X_q|E)$ of a request variableX_q, if a certain evidence E has been observed in the Bayesian network.

As a result of the definition of a conditional probability, the aposteriori probability is

(3) $$P(X_q|E) = \frac{P(X_q, E)}{P(E)} = \frac{\sum_{X \setminus \{X_q, X_E\}} P(X)}{\sum_{X \setminus X_E} P(X)}$$

with X_E designating the quantity of the observed variables.

To overcome the time complexity, the different methods of exact interference calculation use the general principle of dynamic programming.

As part of this exemplary embodiment a simple interference algorithm, of "bucket elimination" [6], is used.

The basic idea with this interference algorithm consists of eliminating variables one after the other in accordance with an order of elimination p by summation.

In this way $P(X_q|E)$ can be efficiently calculated within a perceivable time.

Interventional modelling by setting the evidence

With the interventional modelling approach the effect of specific observation on the behavior of the Bayesian network using a combination of probabilistic interference and data sampling is estimated.

In accordance with **Fig. 3** the Bayesian network can be viewed as a kind of black box 300, with the input being given by a set of observations E 310 and the corresponding list of observed variables X_E 320.

The output, which is given by the data set $D_{B|E}$ 330 is created using the method previously explained in association with **Fig. 2**.

In addition the empirical evidence is to be taken into account.

Consequently each state of X_i is selected with probability $P(state|Pa_i,E)$, which is calculated by means of probabilistic interference.

With the procedure described in accordance with **Fig. 3** different data sets can now be created which reflect the effect of the different observations.

If, as described below, biological effects are analyzed, this means that through this method of operation in accordance with **Fig. 3** artificial microarray data can be created which reflects the probability distribution of a certain data set if specific observations are given.

If the artificially created data from a known origin is compared for example with a cancer-specific set of measurement data, those genes can be determined which, when they are fixed at a certain expression level, will influence the model so that these two microarray data sets, the artificial and the known, exhibit the same characteristics.

Statistical comparison of data sets

In order to estimate the quality of the influence of the evidence I on the behavior of the Bayesian network I, the created data set $D_{B|E}$ is compared with a set of data sets I of known states S.

It is assumed that D describes the effect of different types of cancer. In accordance with the embodiment the behavior of evidence E relating to a specific type of cancer S can now be described.

By using a measure of distance the change a of the correlation

between $D_{B|E}$ and Ds as a result of E can be estimated:

(4) $$a(E) = \frac{d(D_{B|E}, D_S)}{d(D_B, D_S)}$$

with the distance between the two data sets having been standardized with the aid of the distance between D_B, which was taken from B without evidence, and Ds.

As a result, in accordance with the embodiment, the influence of an observed evidence is measurable, e.g. the expression state of a specific gene on a behavior of the model characteristic for cancer.

Secondly the probability can be calculated of B creating a data set DBIE which is equal to Ds for a given E.

For this purpose an estimate is made of how many samples d^1 of $D_{B|E}$ lie closest to Ds in that the distance between each sample and each data set is calculated by D.

The aposteriori probability $P(S|E)$ of the occurence of the cancer type S for given evidence E is thus obtained:

(5) $$P(S|E) = \frac{N_{Es}}{N}$$

with N_{es} being a number of samples of $D_{B|E}$, which is statistically closest to the data set DS, and with N being the total number of samples of $D_{B|E}$.

As already pointed out above, empirical research deals with the relationship between cause and effect, in that it draws conclusions about the underlying cause from experimental observation.

With the Bayesian Inverse Modelling approach in accordance with the exemplary embodiment an underlying cause is estimated by first creating an effect which stems from a known observation.

After this inverse step this effect is compared with effects which are well-defined but for which the cause is unknown.

The potential cause of the best-match effect is then given by the observation which gives rise to the created effect.

The ALL microarray data set of Yeoh et al. [4]

The data which is used for the analysis in accordance with the exemplary embodiment consists of 327 samples of various subtypes of pediatric acute lymphoblastic leukemia (ALL).

The data set was assembled by Yeoh and his colleagues at the St. Jude Children's Research Hospital [4].

ALL is a heterogeneous illness which includes different subtypes, including both T-cell type leukemia and B-cell type leukemia, which differ as regards their reaction to a medical treatment.

Apart from T-ALL, of which the cause is not clearly known, each B-cell subtype can be traced back to a specific genetic modification, e.g. to genetic translocations t(9;22) [BCR-ABL], t(1;19) [E2A-PBX1], t(12;21) [TEL-AML1], t(4;11) [MLL] or to a hyperdiploid karyotype [> 50 chromosomes].

No wonder then that the gene expression patterns of the different subtypes differ very markedly from one another.

Furthermore microarray data exhibits one more clear expression profile which points to the existence of a further ALL subtype in addition to the 6 known.

It should be pointed out that Yeoh et al. [4] are working on a robust classification for classifying the subtypes using a support vector machine with a set of 271 discriminating genes.

Results

Learnt structure

For analysis in accordance with the exemplary embodiment the reduced data set of 271 genes and 327 samples of different ALL subtypes [4], as described above, is used.

To perform the learning process of a multivariate model the data set in the values has been divided up into the discrete value "under-expressed", "expressed normally" and "over-expressed".

The learnt structure shows scale-free characteristic values, a feature which is typical of biological networks, such as for metabolic networks or signaling networks.

Such networks are characterized by a power distribution of the ranges of a node which is defined as the number of connections to other nodes.

These nodes have a strong influence on the dynamics and robustness of scale-free networks, and of many of these strongly connected genes in our model it is actually known that they play a role in the ocogenesis or in the critical processes associated with the development of cancer, e.g. DNA repair.

First a data set of 300 samples is now created from the model in order to estimate the statistics which are defined by the set of the conditional probabilities.

Fig. 4 shows that data obtained by taking samples (**Fig. 4b**)

shows subtype characteristic expression patterns, as is also the case in the original data set (**Fig. 4a**).

The patterns of a number of subtypes such as E2A-PBX1 or T-ALL, are reproduced very well whereas others are generated less well, e.g. the pattern of the subtype MLL, or are missed completely such as for example BCR-ABL.

Modelling of leukaemia subtypes by intervention

The learnt Bayesian network is the basic starting point for the exemplary embodiment for the approach adopted of using inverse modelling to find those genes which, when fixed at a specific expression level, influence the model such that the generated artificial microarray data set exhibits specific characteristics.

As described above, the probability $P(C|E)$ of creation of specific cancer subtype C is estimated if a certain observation E is given, in this case the expression state of a specific gene $P(C|\ Gen_i=\text{state})$.

By contrast with Yeoh, not only the presence of a specific cancer subtype is predicted, but genetic mechanisms which lead to its creation.

A high probability indicates that the fixed gene is a potential cause for the subtype-specific expression behavior of the gene in question, which in its turn can be the underlying cause of a specific cancerous appearance.

7 reference data sets are used for the comparison, with each of these having been obtained in conjunction with a specific ALL subtype.

FIG.4a shows that the original microarray data set is clearly subdivided into 7 clusters (accumulations of points) with

different sample extents.

Each of these clusters represents the expression pattern of 271 genes if a specific subtype of leukaemia is given, and has been used to to measure the influence of an evidence for the occurrence of these different ALL subtypes.

In a first step each gene is fixed for any one of its expression values, with all these conditions being used to to generate a data set of 300 samples (**Fig.4b**).

Subsequently all this data is compared with the 7 reference data sets, as explained previously.

In **Fig. 5** the probability of each subtype, under the condition that a gene is overexpressed, is shown on a graph for 271 genes.

Fig.5 shows that a small number of genes exist which are very likely to trigger a specific ALL subtype if they are strongly active.

To verify these results the molecular function of specific genes and their role in biological processes, especially as regards pathogenesis, is examined in more detail below.

Biological insights

These are obtained by examining in greater detail the genes which are very probably the cause of a specific subtype as well as significant structure patterns in the learnt network, i.e. dominant genes and their environment.

The learnt Bayesian network (model) results from the microarray data set of different leukaemia subtypes and reflects transcriptional relationships between genes which occur in these malignant cancer cells.

Thus genes which trigger a specific subtype are either potential oncogenes or are regulated by such genes.

The first gene to be analyzed in more detail is the gene PBX1.

If it is overexpressed the learnt Bayesan network creates a data set with 0.96 probability which is characteristic of the subtype E2A-PBX1 of the ALL off B-cell type (see **FIG.5**).

This makes the obvious assumption that a causal relationship between the "overexpression" of this gene and the occurrence of the ALL subtypes E2A-PBX1 is present.

And in actual fact PBX1 s known as a proto ocogene which causes normal blood cells to mutate into malignant ALL cancer cells.

As a result of the chromosome translocation t(1;19) PBX1 merges with the gene E2A and transform into a potent ocogene which causes the leukemia subtype E2A-PBX1.

Since the graph structure of the model (**Fig. 6**) can further be interpreted in a causal manner it provides information about the interaction between potential oncogenes and other genes which in its turn can be interpreted as an oncogene regulation.

Ithe structure of the network (**Fig.6**) is considered, PBX1 represents a dominant gene in that it influences many other genes but is only regulated by one or a few other genes.

In addition, as a result of the conditional probability distribution, the model identifies PBX1 as a transcription activator.

This can also be explained by known biological facts, since PBX1 activates genes which are normally not expressed or are

expressed at a low level.

Patients with a hyperdiploidy of > 50 chromosomes have clones of 51-68 chromosomes. Although high hyperdiploid clones are seldom identical, they tend to exhibit a pattern of the chromosome increase with additional copies of the chromosomes 4, 6, 10, 14, 18 and 21.

Trisomy and Polysomy 21 are non-random anomalies which are frequently to be observed with ALL Their occurrence, even if it is not specific, as well as the increased occurrence of acute leukaemia or in subjects with constitutional Trisomy 21 make it reasonable to assume that the chromosome 21 has a particular role to play in leukemogenesis.

Another disease, Down's Syndrome, is caused by Trisomy 21 and shows an increased occurence of leukemia such as ALL.

As a result the method described makes it possible in this case, in accordance with the exemplary embodiment, to identify genes which to a large extent indicate the hyperdiploid ALL subtype, of which however it is also known that they play a significant role in the occurrence of Down's Syndrome.

The gene SOD1 is located at chromosom e21 and produces an enzyme which converts superoxide-free radicals into hydrogen peroxide. The increased expression at Trisomy 21, which is also to be observed for the microarray samples of patients with hyperdiploid karyotype, can give rise to the brain damage which is to be seen with Down's Syndrome.

The frequency of the occurence of the hyperdiploid ALL also increases in the case in which the gene PSMD10 is overexpressed.

PSMD10 is a regulatory cluster unit of the proteasome 26S for which it has been shown that is operates as a natural

mechanism for the breakdown of protein by regulating the protein metabolism in eukaryotic cells

This is of significance for cancers in humans since the cell cycle, the growth of the tumor and the survival are determined by a great vraiety of intracellular proteins which are regulated by the ubiquitin-dependent proteasome breakdown path which is influenced by PSMD10.

In more recent scientific work it has been verified that this breakdown path is often the object of a deregulation associated with cancer and can be subject to such processes as oncogene transformation, tumor progression, bypassing of the immune system and resistance to medicaments.

Abstract of the exemplary embodiment

The exemplary embodiment described presents a new method by which it is possible to identify genes which are a potential cause of tumorgenesis, by analyzing the relationships between microarray data of leukemia subtypes and a data set, which is the result of taking samples from a learnt Bayesian network.

This method of operation is based on the modelling of a regulator genetic network through a Bayesian network, with genes or their corresponding proteins being symbolized by the nodes of the Bayesian network.

Regulation mechanisms are described by connectors between two nodes, which can be interpreted in a causal manner.

The quality of the regulation is encoded in the conditional probability dsitribution of the gene involved for given regulators of the same.

The understanding of the regulatory genetic network represents an important step along the road to characterizing the genetic

mechanisms underlying complex diseases.

In cancer research, were the identification of genes which suppress growths and tumors plays a key role, the knowledge of new potential oncogenes and their interactions with other molecules is an important contribution to discovering the basic principles which determine why normal cells mutate into malignant cancer cells.

With the procedure described in accordance with the exemplary embodiment, especially with Bayesian Inverse Modelling, it is possible to discover genes with such an oncogene characteristic simply through a statistical analysis of gene expression patterns, which have been measured with the aid of DNA microarrays.

The underlying theoretical probability model which has been used, is a Bayesian network, which encodes the multivariate probability distribution of a set of variables by means of a set of conditional probability distributions.

The statistical dependencies are encoded in a graph structure. In the learning method Bayesian statistics are used to determine the network structure and the corresponding model parameters which best describe the probability distribution contained in the data.

The following publications are cited in this document:

[1] Stetter Martin et al., Large-Scale Computational Modeling of Generic Regulatory Networks, Kluwer Academic Publisher, Netherlands, 2003;

[2] Publication number DE 10159262.0;

[3] F. W. Jensen, F. V. (1996), An introduction to Bayesian networks, UCL Press, London; 178 pages;

[4] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Petal et al. (2002), Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profile. Cancer cell 1:133-143;

[5] D. Heckerman, D. Geiger and D. Chickering (1995), Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning 20:197-243;

[6] R. Dechter (1996), Bucket elimination: A unifying framework for probabilistic inference. In: Uncertainty in Artificial intelligence, UA196:211-219.